# OT-ALD: Aligning Latent Distributions with Optimal Transport for Accelerated Image-to-Image Translation

**Zhanpeng Wang[1], Shuting Cao[1], Yuhang Lu[1], YuhanLi[1], Na Lei[1*], Zhongxuan Luo[1]**

[1]Dalian University of Technology, Dalian, Liaoning, China

## Abstract

The Dual Diffusion Implicit Bridge (DDIB) is an emerging image-to-image (I2I) translation method that preserves cycle consistency while achieving strong flexibility. It links two independently trained diffusion models (DMs) in the source and target domains by first adding noise to a source image to obtain a latent code, then denoising it in the target domain to generate the translated image. However, this method faces two key challenges: (1) low translation efficiency, and (2) translation trajectory deviations caused by mismatched latent distributions. To address these issues, we propose a novel I2I translation framework, OT-ALD, grounded in optimal transport (OT) theory, which retains the strengths of DDIB-based approach. Specifically, we compute an OT map from the latent distribution of the source domain to that of the target domain, and use the mapped distribution as the starting point for the reverse diffusion process in the target domain. Our error analysis confirms that OT-ALD eliminates latent distribution mismatches. Moreover, OT-ALD effectively balances faster image translation with improved image quality. Experiments on four translation tasks across three high-resolution datasets show that OT-ALD improves sampling efficiency by 20.29% and reduces the FID score by 2.6 on average compared to the top-performing baseline models.

## Introduction

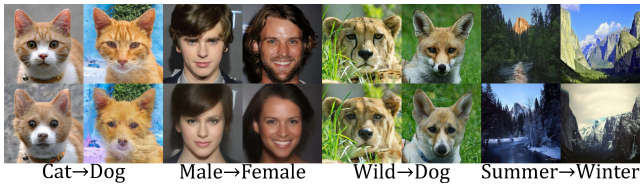
Image-to-image (I2I) translation is a fundamental task in computer vision that involves transforming an image from a source domain to a target domain while preserving its structural and semantic integrity. This technique has broad applications, including image restoration (Zamir et al. 2022), style transfer (Azadi et al. 2018), and image synthesis (Rombach et al. 2022), among others. The rapid advancement of deep generative models has led to significant progress in I2I translation, giving rise to numerous state-of-the-art methods. In particular, generative adversarial networks (GANs) (Zhu et al. 2017; Isola et al. 2017; Yi et al. 2017; Fu et al. 2019; Park et al. 2020) have played a pivotal role in early developments. More recently, diffusion models (DMs) (Sasaki, Willcocks, and Breckon 2021; Su et al. 2022; Zhao et al. 2022; Saharia et al. 2022a; Li et al. 2023; Meng et al. 2021) have gained prominence and become a major focus in contemporary research due to their ability to produce high-quality, diverse outputs.

The contraction property of DMs (Khrulkov et al. 2022; Franzese et al. 2023), as illustrated in Figure 1(b), significantly contributes to their rapid adoption in I2I translation. The core mechanism of DMs involves systematically corrupting a real image by progressively adding noise until it becomes pure Gaussian noise (Ho, Jain, and Abbeel 2020). Subsequently, the model learns to reconstruct the original image from noise through score matching (Song et al. 2020). This reconstruction is facilitated via a reverse process that incrementally removes noise, guiding the image from a random state back to the real data distribution. Mathematically, these forward and reverse processes are described by a pair of stochastic differential equations (SDEs) (Song et al. 2020; Anderson 1982). Currently, diffusion-based I2I translation methods can be roughly categorized into two types: (1) Utilizing the source image to guide the generation trajectory of the target image (Zhao et al. 2022; Choi et al. 2021; Meng et al. 2021; Sun et al. 2023; Li and Ma 2023; Kim, Kwon, and Ye 2022; Tumanyan et al. 2023; Seo et al. 2023; Yang et al. 2023), ensuring that the output evolves toward the desired target. However, this approach relies on joint training on both the source and target domains, necessitating simultaneous access to both datasets. This requirement limits data separation, undermines privacy protection, and reduces the overall flexibility. (2) Interconnecting two independently trained DMs in the source and target domains (Su et al. 2022; Zhang et al. 2024; Bourou et al. 2024; Hur et al. 2024; Yin et al. 2024; Mancusi et al. 2024), a technique referred to as the Dual Diffusion Implicit Bridges (DDIB)-based method. Firstly, the source domain DM adds noise to the source image, transforming it into a latent code. The target domain DM then progressively denoises the latent code to generate the target image. This method enables I2I translation while ensuring privacy protection and high flexibility.

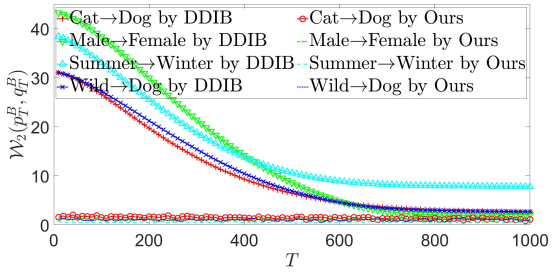
The DDIB-based method is not without its limitations. First, employing two DMs leads to increased inference time (Su et al. 2022). Second, in practical applications, DMs must terminate after a finite number of time steps, meaning the termination distributions (latent code distributions) of the source and target domain DMs do not perfectly align with the standard Gaussian distribution. This misalignment intro-

*Corresponding author. Email: nalei@dlut.edu.cn

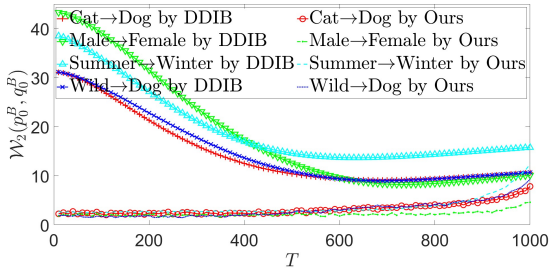
duces a theoretical gap, which has been shown to cause deviations in the image translation trajectory within the target domain. As a result, even with high-precision score matching training for both DMs, the translated images may fail to fully conform to the intended target distribution (see Theorem 1). Furthermore, the cycle consistency of the DDIB-based method may also be compromised.




(a) The translation effect of the proposed model on multiple high-resolution image tasks




(b) Wasserstein difference in latent distributions




(c) Wasserstein difference in target distributions

Figure 1: (a) Top-row images are sources and bottom-row images are corresponding outputs generated by our model. (b) Contraction property of DMs. For initial distributions from different domains, DMs can exponentially shrink Wasserstein distance between their latent distributions through progressive noise injection. $T$ is the number of diffusion steps, and $q_T^B$ is initial distribution for the reverse process of $\mathrm{DM}^B$. In DDIB-based methods, $q_T^B = p_T^A$, while OT-ALD aligns latent distributions via OT map $M_{ot,T}^{A\to B}$ to ensure $q_T^B = M_{ot,T}^{A\to B}(p_T^A)$. (c) Latent alignment impacts how closely the translated distribution approximates the ground truth (Theorem 1). DDIB requires longer diffusion to compensate for misalignment, reducing efficiency.

To address the abovementioned issues, we propose to eliminate this gap by aligning the latent code distribution using an optimal transport (OT) map (see Theorem 2). The main contributions of this paper are summarized as follows:

- We demonstrate that existing DDIB-based I2I translation methods suffer from a theoretical gap caused by a mismatch between latent distributions. As a result, even with perfectly accurate score matching, the translation trajectory may still deviate from the intended target distribution.
- Building on OT, we propose a novel I2I translation framework, OT-ALD, which theoretically aligns the latent distributions. Our theoretical analysis and experimental results show that OT-ALD inherits cycle consistency and strong flexibility of the DDIB-based approach.
- OT-ALD can strike a balance between accelerating I2I translation and enhancing the quality of image generation. Experiments on four tasks across three high-resolution datasets show that compared with the best baseline model, OT-ALD achieves an average 20.29% improvement in sampling efficiency and 2.6 reduction in FID.

## Preliminaries and Related Works

### Image Synthesis Through the DMs

The DMs include forward and reverse processes. Given drift coefficient $\boldsymbol{f}(\boldsymbol{x},t):\mathbb{R}^n\times[0,T]\to\mathbb{R}^n$ and diffusion term $g(t):[0,T]\to\mathbb{R}_{>0}$, the internal mechanism of the forward process can be elucidated by SDE (Song et al. 2020)

$$d\boldsymbol{x}=\boldsymbol{f}(\boldsymbol{x},t)\,dt+g(t)\,d\boldsymbol{W}_t,\boldsymbol{x}(0)\sim p_0(\boldsymbol{x}), \quad (1)$$

where $p_0(\boldsymbol{x})$ is the image distribution and $\boldsymbol{W}_t$ is the $n$-dimensional standard Wiener process. The marginal distribution of the solution to (1), $p_t(\boldsymbol{x})$, satisfies the Fokker-Planck equation (FPE)

$$\frac{\partial p_t(\boldsymbol{x})}{\partial t}+\nabla\cdot(\boldsymbol{f}(\boldsymbol{x},t)\,p_t(\boldsymbol{x}))-\frac{g(t)^2}{2}\Delta p_t(\boldsymbol{x})=0. \quad (2)$$

According to (2), there exists a deterministic process, referred to as the probability flow ordinary differential equation (PF-ODE) (Song et al. 2020), which shares the same marginal probability density $p_t$ as (1)

$$d\boldsymbol{x}=(\boldsymbol{f}(\boldsymbol{x},t)-\frac{g(t)^2}{2}\nabla\log p_t(\boldsymbol{x}))dt,\boldsymbol{x}(0)\sim p_0(\boldsymbol{x}).$$

Similarly, the reverse process can be expressed as reverse-time SDE (Anderson 1982) with $q_T$ as the initial condition

$$d\boldsymbol{x}=(\boldsymbol{f}(\boldsymbol{x},t)-g(t)^2\nabla\log p_t(\boldsymbol{x}))dt+g(t)\,d\bar{\boldsymbol{W}}_t, \quad (3)$$

where $\boldsymbol{x}(T)\sim q_T(\boldsymbol{x})$ and $\bar{\boldsymbol{W}}_t$ is the $n$-dimensional standard Wiener process with backward time. In practice, $\nabla\log p_t(\boldsymbol{x})$ is usually approximated by a score network $\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{x},t):\mathbb{R}^n\times[0,T]\to\mathbb{R}^n$ with the weighted mean square error (MSE) loss function (Song et al. 2021)

$$\mathcal{J}_{SM}:=\frac{1}{2}\int_0^T\phi(t)\,\mathbb{E}_{p_t}[\|\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{x},t)-\nabla\log p_t(\boldsymbol{x})\|_2^2]dt,$$

here $\phi(t):[0,T]\to\mathbb{R}_{>0}$ is a positive weighting function. By replacing $\nabla\log p_t(\boldsymbol{x})$ with $\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{x},t)$ in the reverse process, we get an approximate reverse-time SDE and FPE. Ultimately, the approximate image distribution $q_0$ is obtained from $q_T$. The solution of DMs can be formalized as

$$\begin{aligned}&Solver_\eta(\boldsymbol{x}_{t_0},\boldsymbol{S}_{\boldsymbol{\theta}},t_0,t_1)\\=&\boldsymbol{x}_{t_0}+\int_{t_0}^{t_1}\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x},t)\,dt+\eta\int_{t_0}^{t_1}g(t)d\boldsymbol{W}_t,\end{aligned} \quad (4)$$

where $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x},t)=\boldsymbol{f}(\boldsymbol{x},t)-\frac{1}{2}(1-\eta^2)g(t)^2\boldsymbol{S}_{\boldsymbol{\theta}}(\boldsymbol{x},t)$ and $\eta\in[0,1]$ denotes noise scaling factor.

## Optimal Transport Theory

We begin by introducing the classical geometric variational approach to semi-discrete OT. As the number of target samples increases, the discrete solution progressively converges to its continuous counterpart. Suppose the source measure $\mu$ is defined on a convex domain $\Omega \subset \mathbb{R}^n$, and the target domain is a discrete set $\mathcal{Y} = \{\boldsymbol{y}_i\}_{i\in\mathcal{I}} \subset \mathbb{R}^n$. The target measure is a Dirac measure $\nu = \sum_{i\in\mathcal{I}} \nu_i \delta(\boldsymbol{y} - \boldsymbol{y}_i)$, with the total mass matched as $\mu(\Omega) = \sum_{i\in\mathcal{I}} \nu_i$. Under a transport map $M : \Omega \to \mathcal{Y}$, the domain $\Omega$ is partitioned into cells $W_i$ such that each point $\boldsymbol{x} \in W_i$ is mapped to $\boldsymbol{y}_i$, i.e., $M(\boldsymbol{x}) = \boldsymbol{y}_i$. The map $M$ is measure-preserving (denoted $M_{\#}\mu = \nu$) if $\mu(W_i) = \nu_i, \forall i \in \mathcal{I}$. Let $c : \Omega \times \mathcal{Y} \to \mathbb{R}$ be cost function, where $c(\boldsymbol{x}, \boldsymbol{y})$ denotes the cost of transporting unit mass from $\boldsymbol{x}$ to $\boldsymbol{y}$. The total transport cost is given by

$$\int_\Omega c(\boldsymbol{x}, M(\boldsymbol{x}))d\mu(\boldsymbol{x}) = \sum_{i\in\mathcal{I}} \int_{W_i} c\left(\boldsymbol{x}, \boldsymbol{y}_i\right) d\mu(\boldsymbol{x}). \quad (5)$$

The OT map $M_{ot}$ is a measure-preserving map that minimizes the total cost in (5),

$$M_{ot} := \arg\min_{M_{\#}\mu=\nu} \int_\Omega c(\boldsymbol{x}, M(\boldsymbol{x}))d\mu(\boldsymbol{x}). \quad (6)$$

Specifically, when the cost function is $c(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$, there exists a convex function $u : \Omega \to \mathbb{R}$ such that its gradient $\nabla u$ uniquely solves (6) (Brenier 1991), i.e., $M_{ot} = \nabla u$. This shows that the OT map is the gradient of Brenier's potential. As noted in (Lei et al. 2020), $u$ can be represented as the upper envelope of hyperplanes $\pi_{\boldsymbol{h},i}(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{y}_i\rangle + h_i$, and is uniquely parameterized (up to an additive constant) by a height vector $\boldsymbol{h} = (h_1, h_2, \ldots, h_{|\mathcal{I}|})^T$. The resulting potential $u_{\boldsymbol{h}}(\boldsymbol{x}) = \max_{i\in\mathcal{I}}\{\pi_{\boldsymbol{h},i}(\boldsymbol{x})\}$ satisfies the projected volume under each support plane matches the prescribed mass $\nu_i$. To obtain $u_{\boldsymbol{h}}$, it suffices to optimize the height vector $\boldsymbol{h}$ by minimizing the following convex energy:

$$E(\boldsymbol{h}) = \int_{\boldsymbol{0}}^{\boldsymbol{h}} \sum_{i\in\mathcal{I}} w_i(\gamma)d\gamma_i - \sum_{i\in\mathcal{I}} h_i\nu_i, \quad (7)$$

where $\omega_i(\gamma)$ is the $\mu$-volume of $W_i(\gamma)$. For more theoretical details, we refer to Reference (Gu et al. 2013).

## Related Works

This section presents a curated overview of relevant literature, with additional references available in *Appendix A* and in the review articles (Huang et al. 2024).

**Conditional guided and controlled translation.** To enhance control over image generation, researchers have proposed multimodal conditional strategies using text, exemplars, or semantic labels. Kim et al. (Kim, Kwon, and Ye 2022) applied DMs with CLIP loss for text-guided editing, while Meng et al. (Meng et al. 2021) synthesized realistic images by adding noise to images and denoising via reverse-time SDEs. Tumanyan et al. (Tumanyan et al. 2023) enabled fine-grained control by manipulating internal features of pre-trained models. Seo et al. (Seo et al. 2023) improved exemplar-guided accuracy via alternating cross-domain matching and latent diffusion. Cheng et al. (Cheng et al. 2023) addressed single-example translation through content-concept inversion and fusion. Classifier-free guidance was adopted in (Yang et al. 2023) for controlled editing. Kwon et al. (Kwon and Ye 2022) performed text- and image-guided style transfer by aligning attention keys and class tokens, while Shi et al. (Shi et al. 2024) introduced latent optimization with reference latent control for point-interactive editing.

**Efficient and lightweight translation.** To mitigate the high computational cost of DMs, various optimization strategies have been explored. Song et al. (Song, Meng, and Ermon 2021) introduced non-Markovian sampling to reduce generation steps, while Luo et al. (Luo et al. 2023) achieved single-step inference via consistency distillation. Xia et al. (Xia et al. 2024a) proposed compact prior networks and dynamic transformers, and Jiang et al. (Jiang et al. 2024; Xia et al. 2024b) optimized time-step usage to accelerate translation. Parmar et al. (Parmar et al. 2024) combined single-step DMs with adversarial learning for fast, high-quality synthesis. Lee et al. (Lee, Jeong, and Sohn 2024) employed Brownian bridge strategies for better stability and efficiency. Additionally, Lee et al. (Lee, Kang, and Han 2024) improved text-driven I2I translation by interpolating prompts to refine noise prediction in pre-trained DMs.

# Methodology of OT-ALD

We use superscripts $^A$ and $^B$ to indicate their association with the source and target domains, respectively. Figure 2 illustrates the essential differences between OT-ALD and DDIB-based methods (Su et al. 2022; Zhang et al. 2024; Bourou et al. 2024; Hur et al. 2024; Yin et al. 2024; Mancusi et al. 2024). Given an initial image distribution $p_0^A$ in the source domain, DDIB-based methods advocate using the forward process of $\mathrm{DM}^A$ to corrode $p_0^A$ into a latent code distribution $p_T^A$. The reverse process of $\mathrm{DM}^B$ then reconstructs the target image distribution $p_0^B$ from $p_T^A$. According to the contraction property of SDEs, as illustrated in Figure 1(b), the Wasserstein distance $\mathcal{W}_2(p_T^A, p_T^B)$ converges exponentially to zero as $T \to +\infty$ (Khrulkov et al. 2022; Franzese et al. 2023). However, in practical scenarios with finite $T$, we have $\mathcal{W}_2(p_T^A, p_T^B) \neq 0$, resulting in a mismatch between latent code distributions. Here we present Theorem 1, which demonstrates that this mismatch leads to deviations in the translation trajectory within the target domain, ultimately preventing the attainment of the desired target distribution. The proof can be found in *Appendix D.1*.

**Theorem 1.** *If $q_T^B = p_T^A$, we denote $q_0^B$ as the distribution generated by the DM$^B$, then $\mathcal{W}_2(p_0^B, q_0^B)$ can be estimated as follows*

$$\bar{I}^B(T)\,\mathcal{W}_2(p_T^A, p_T^B) \leq \mathcal{W}_2(p_0^B, q_0^B) \leq I^B(T)\,\mathcal{W}_2(p_T^A, p_T^B),$$

*where $I^B(T) = \exp\left(\int_0^T (L_f^B(t) + \frac{g^B(t)^2}{2} L_{\boldsymbol{S_\theta}}^B(t))dt\right)$ and $\bar{I}^B(T) = \exp\left(\frac{1}{2}\int_0^T L_f^B(t)\,dt\right)$, $L_f^B(t)$ and $L_{\boldsymbol{S_\theta}}^B(t)$ mean the continuous Lipschitz constant shown in Appendix B.*

Theorem 1 states that for DDIB-based methods whenever $\mathcal{W}_2(p_T^A, p_T^B) \neq 0$ holds, $\mathcal{W}_2(p_0^B, q_0^B) \neq 0$ must follow. To address this theoretical gap, we compute the OT map

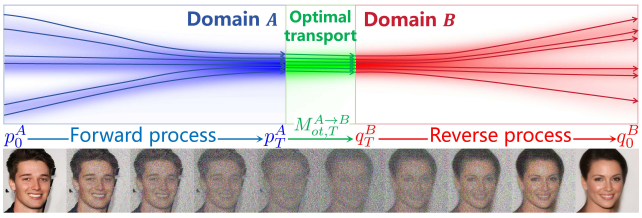



Figure 2: The framework of OT-ALD. During training, two DMs are independently trained in domains $A$ and $B$, followed by computation of the OT map $M_{ot,T}^{A\to B}$ from $p_T^A$ to $p_T^B$. In translation, the source distribution $p_0^A$ is diffused to $p_T^A$, which is then mapped via $M_{ot,T}^{A\to B}$ to serve as the initial distribution for the reverse process in domain $B$, yielding the translated distribution $q_0^B$. In contrast, DDIB-based methods skip OT alignment and directly use $p_T^A$, leading to latent distribution mismatch. As shown in Theorem 1, this mismatch introduces a theoretical gap that affects translation accuracy.

$M_{ot,T}^{A\to B}$ to align the source domain latent code distribution $p_T^A$ with the target domain latent code distribution $p_T^B$ (See Algorithm 1). We then use $M_{ot,T}^{A\to B}(p_T^A)$ as the initial distribution for the reverse process of DM$^B$, effectively eliminating the discrepancy caused by latent code distribution mismatch. Furthermore, we provide an error estimation in Theorem 2. The proof can be found in *Appendix D.2*.

**Theorem 2.** *The error upper bound between the distribution obtained by OT-ALD and the ground truth is*

$$\mathcal{W}_2(p_0^B, q_0^B) \le \sqrt{2T}(\mathcal{J}_{SM}^B)^{\frac{1}{2}} + KI^B(T)\|u_T^{A\to B} - u_{\boldsymbol{h}}\|_\infty^{\frac{1}{2}},$$

*where $K$ is a positive constant and $u_T^{A\to B}$ is the true Brenier function between $p_T^A$ and $p_T^B$.*

It is worth noting that the noise scaling factor $\eta \in [0,1]$ in (4) determines whether the diffusion process is Markovian or non-Markovian by controlling the noise added during sampling (Song, Meng, and Ermon 2021). When $\eta \to 1$, our model generates images with greater diversity. Conversely, as $\eta \to 0$, it produces images more quickly and with higher determinism (See Figure 3(a)).

**Complexity analysis.** In Algorithm 1, the complexity of determining the cell to which each point in set $\mathcal{X}_T^A$ belongs is $\mathcal{O}(|\mathcal{K}||\mathcal{I}|)$. Subsequently, the complexity of calculating the measure of a cell based on the points belonging to it is $\mathcal{O}(|\mathcal{K}||\mathcal{I}|)$. Assuming the number of iterations in Algorithm 1 is $N$, the complexity introduced by the Adam algorithm is $\mathcal{O}(N|\mathcal{I}|)$. Therefore, the total complexity of computing OT is $\mathcal{O}((2|\mathcal{K}| + N)|\mathcal{I}|)$.

Therefore, the OT-ALD consists of three stages (see Algorithm 2): (1) Compute latent code $\boldsymbol{x}_T^A$ by applying the forward process of DM$^A$ to degrade source image $\boldsymbol{x}_0^A$. (2) Obtain latent code $\boldsymbol{x}_T^B$ in the target domain corresponding to $\boldsymbol{x}_T^A$ under OT map $M_{ot,T}^{A\to B}$. (3) Generate target image $\boldsymbol{x}_0^B$ by applying the reverse process of DM$^B$ to reconstruct $\boldsymbol{x}_T^B$.

Building on its ability to eliminate latent code distribution mismatch, OT-ALD can be naturally extended to accelerated I2I translation by reducing the diffusion termination time $T$.

Algorithm 1: Computing OT between latent distributions.

**Input**: Source images $\mathcal{X}_0^A = \{\boldsymbol{x}_{0,k}^A\}_{k\in\mathcal{K}} \sim p_0^A$, target images $\mathcal{X}_0^B = \{\boldsymbol{x}_{0,i}^B\}_{i\in\mathcal{I}} \sim p_0^B$, trained $\boldsymbol{S}_{\boldsymbol{\theta}}^A$ and $\boldsymbol{S}_{\boldsymbol{\theta}}^B$, diffusion termination time $T$.
**Output**: OT map $M_{ot,T}^{A\to B}$.

$\mathcal{X}_T^A \longleftarrow Solver_\eta(\mathcal{X}_0^A, \boldsymbol{S}_{\boldsymbol{\theta}}^A, 0, T)$.
$\mathcal{X}_T^B \longleftarrow Solver_\eta(\mathcal{X}_0^B, \boldsymbol{S}_{\boldsymbol{\theta}}^B, 0, T)$.
$\boldsymbol{h} \longleftarrow \mathbf{0}$.
**repeat**
  Calculate $\nabla E(\boldsymbol{h}) = (w_i(\boldsymbol{h}) - \nu_i)^T$.
  $\nabla E(\boldsymbol{h}) = \nabla E(\boldsymbol{h}) - mean(\nabla E(\boldsymbol{h}))$.
  Update $\boldsymbol{h}$ by Adam algorithm ($\beta_1 = 0.9$, $\beta_2 = 0.5$).
**until** Converge
$M_{ot,T}^{A\to B} \longleftarrow \nabla u_{\boldsymbol{h}}$, $u_{\boldsymbol{h}}(\cdot) = \max_i \langle \cdot, \boldsymbol{x}_{T,i}^B \rangle + h_i$.

**Return**: $M_{ot,T}^{A\to B}$.

Algorithm 2: I2I translation process of OT-ALD.

**Input**: Source image $\boldsymbol{x}_0^A \sim p_0^A$, computed OT map $M_{ot,T}^{A\to B}$, trained score networks $\boldsymbol{S}_{\boldsymbol{\theta}}^A(\boldsymbol{x}^A, t)$ and $\boldsymbol{S}_{\boldsymbol{\theta}}^B(\boldsymbol{x}^B, t)$, noise scaling factor $\eta$.
**Output**: Target image $\boldsymbol{x}_0^B$.

$\boldsymbol{x}_T^A \longleftarrow Solver_\eta(\boldsymbol{x}_0^A, \boldsymbol{S}_{\boldsymbol{\theta}}^A, 0, T)$.
$\boldsymbol{x}_T^B \longleftarrow M_{ot,T}^{A\to B}(\boldsymbol{x}_T^A)$.
$\boldsymbol{x}_0^B \longleftarrow Solver_\eta(\boldsymbol{x}_T^B, \boldsymbol{S}_{\boldsymbol{\theta}}^B, T, 0)$.

**Return**: $\boldsymbol{x}_0^B$.

Figure 3(b) highlights that OT-ALD with a smaller $T$ better preserves the details and structural integrity of the source image, while a larger $T$ leverages the strengths of DM to enhance the diversity of the translated image. Therefore, an appropriate value of $T$ must be selected to strike a balance between computational efficiency and output quality.

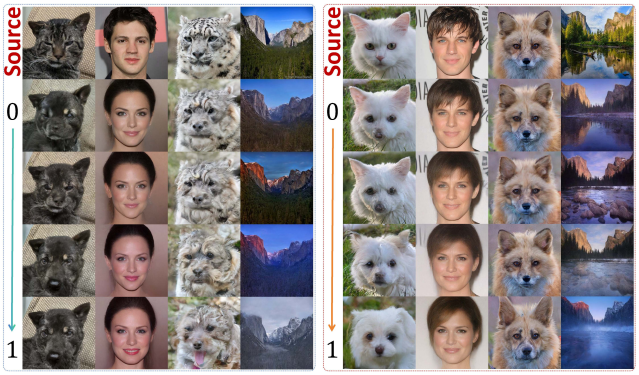
Theoretically, OT-ALD satisfies cycle consistency, which we explain from two different views: sample-level (Theorem 3) and distribution-level (Theorem 4). Notably, sample-level cycle consistency holds only when $\eta = 0$ in Algorithm 2. However, once stochasticity is introduced, this property no longer holds. In contrast, distribution-level cycle consistency remains valid for $\forall \eta \in [0,1]$ under the control of FPE (2).

**Theorem 3** (Sample cycle consistency)**.** *Disregarding the score matching error, discretization error of the $Solver_\eta$ and computational error of the OT map, $\eta = 0$. For any $T \ge 0$ and source image $\boldsymbol{x}_0^A$, we consider following cyclic process*

$$\boldsymbol{x}_T^A = Solver_\eta(\boldsymbol{x}_0^A, \boldsymbol{S}_{\boldsymbol{\theta}}^A, 0, T), \boldsymbol{x}_T^B = M_{ot,T}^{A\to B}(\boldsymbol{x}_T^A),$$
$$\boldsymbol{x}_0^B = Solver_\eta(\boldsymbol{x}_T^B, \boldsymbol{S}_{\boldsymbol{\theta}}^B, T, 0), \boldsymbol{x}_T'^B = Solver_\eta(\boldsymbol{x}_T^B, \boldsymbol{S}_{\boldsymbol{\theta}}^B, 0, T),$$
$$\boldsymbol{x}_T'^A = M_{ot,T}^{B\to A}(\boldsymbol{x}_T'^B), \boldsymbol{x}_0'^A = Solver_\eta(\boldsymbol{x}_T'^A, \boldsymbol{S}_{\boldsymbol{\theta}}^A, T, 0),$$

*where $M_{ot,T}^{B\to A} = (M_{ot,T}^{A\to B})^{-1}$, then $\|\boldsymbol{x}_0^A - \boldsymbol{x}_0'^A\|_2 = 0$ holds. See Appendix D.3 for the proof detail.*

**Theorem 4** (Distributional cycle consistency)**.** *Under the same setting as Theorem 3, we apply the proposed method to translate the source distribution $p_0^A$ into the target distribution. Subsequently, the approximate source distribution*




(a) $T = 500, \eta : 0 \to 1$. (b) $\eta = 0.4, T/1000 : 0 \to 1$.

Figure 3: The smaller the adopted $\eta$ and $T$, the higher the fidelity of OT-ALD to the source image is maintained; Conversely , larger values of $\eta$ and $T$ indicate that OT-ALD can achieve greater diversity in I2I translation tasks.

$p_0'^A$ *derived from the reverse translation action on the target distribution. Then we have* $\mathcal{W}_2(p_0^A, p_0'^A) = 0$ *for* $\forall \eta \in [0, 1]$. *See Appendix D.4 for the proof detail.*

Sample consistency forms the basis of distribution consistency—when all individuals satisfy cycle consistency, the distribution naturally aligns. However, distribution consistency does not require strict reversibility of individual samples, making it more flexible and tolerant of individual differences to improve overall distribution alignment.


Figure 4: The flexibility tests of OT-ALD on the trained DMs and OT map. Top-row images are sources; bottom-row images are the corresponding outputs generated by OT-ALD.

Moreover, OT-ALD is based on independently trained DMs and OT map, thereby inheriting the strong flexibility advantages of DDIB-based methods. To validate this property, we recombine trained DMs in a manner different from the four main experiments. As shown in Figure 4, OT-ALD still achieves high-quality translation results under this setting. Notably, as shown in the "Wild→Wild" result, when the target image dataset is the same as source image dataset, our model can achieve the effect of detail enhancement.

# Experiments

## Implementation Details

**Datasets.** We evaluate four I2I translation tasks on three public datasets. (1) AFHQ (Choi et al. 2020) is a 512×512 animal face dataset containing approximately 5000 images per class, used for Cat→Dog and Wild→Dog translation. (2) Summer2Winter dataset (Zhu et al. 2017) consists of 256×256 seasonal landscape images under varying weather and lighting conditions, used for Summer→Winter translation. (3) CelebA-HQ dataset (Karras et al. 2017) includes 30000 facial images (resized to 512×512) with diverse attributes and poses, used for Male→Female translation.

**Baseline models.** In this work, we compare our model with the following methods: CUT (Park et al. 2020), EGSDE (Zhao et al. 2022), ILVR (Choi et al. 2021), SDEdit (Meng et al. 2021), StarGANv2 (Choi et al. 2020), SDDM (Sun et al. 2023), InjectDiff (Li and Ma 2023), CycleGAN (Zhu et al. 2017), QS-Attn (Hu et al. 2022), DDIB (Su et al. 2022), InstPix2Pix (Brooks, Holynski, and Efros 2023), Pix2Pix-Zero (Parmar et al. 2023), P2P (Hertz et al. 2022), UNSB (Kim et al. 2023), GcGAN (Fu et al. 2019), DistanceGAN (Benaim and Wolf 2017), Plug&Play (Tumanyan et al. 2023), CycleDiff (Wu and De la Torre 2023), Turbo (Parmar et al. 2024), U-GAT-IT (Kim et al. 2019), NICE-GAN (Chen et al. 2020), DDIB (Su et al. 2022; Zhang et al. 2024; Bourou et al. 2024; Hur et al. 2024; Yin et al. 2024; Mancusi et al. 2024), DMT (Xia et al. 2024b).

**Evaluation metrics.** For the translated images, we evaluate realism using Fréchet Inception Distance (FID) (Heusel et al. 2017), Kernel Inception Distance (KID) (Bińkowski et al. 2018) and Fool Rates (Saharia et al. 2022b). To measure faithfulness, we employ $L_2$ distance, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) (Wang et al. 2004), and DINO-Struct-Dist (Tumanyan et al. 2022). Additionally, the Wasserstein distance $\mathcal{W}_2(p_T^B, q_T^B)$ quantifies the latent code distribution matching error of the proposed method and DDIB-based approach. Finally, we assess computational efficiency by comparing translation time (s/image) across different models.

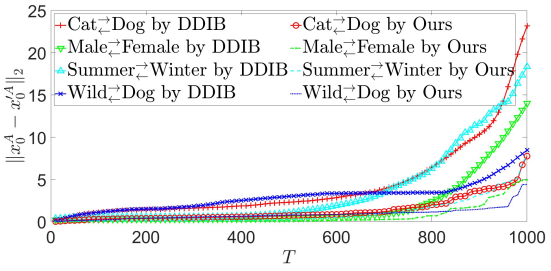
## Mismatch Within Latent Code Distributions and Cycle Consistency

For the four I2I translation tasks, Figure 1(b) shows the latent distribution mismatch for our method and the DDIB-based approach. The results confirm that introducing the OT map effectively eliminates this discrepancy, yielding a target distribution closer to the ground truth (see Figure 1(c), Theorems 1–2). We further analyze the effect of varying diffusion termination time $T$, showing that our method is less sensitive to $T$ than DDIB, which requires a longer diffusion process to compensate for the mismatch—at the cost of significantly lower sampling efficiency.

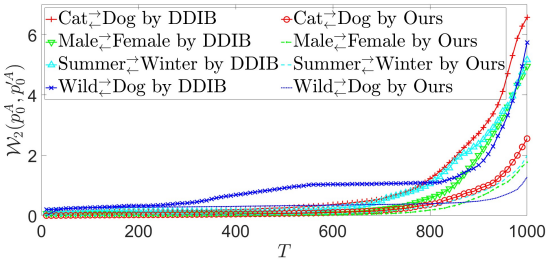
Subsequently, we evaluate the cycle consistency of our method and DDIB-based approach in Figures 5(a) and 5(b). Although both methods theoretically satisfy sample-level and distribution-level cycle consistency (Theorems 3-4), the large number of diffusion steps in the DDIB-based method leads to accumulated computational errors, resulting in inferior cycle consistency compared to our model.

## Image Realism and Faithfulness

Table 1 reports quantitative results of OT-ALD ($T = 500$, $\eta = 0.2$) and baseline models across four I2I translation




(a) Sample cycle consistency




(b) Distribution cycle consistency

Figure 5: Comparison of cycle consistency between OT-ALD and DDIB-based methods.

tasks. Compared to other methods, OT-ALD not only preserves the structural integrity of the source images but also achieves the highest output quality. Across these four tasks,

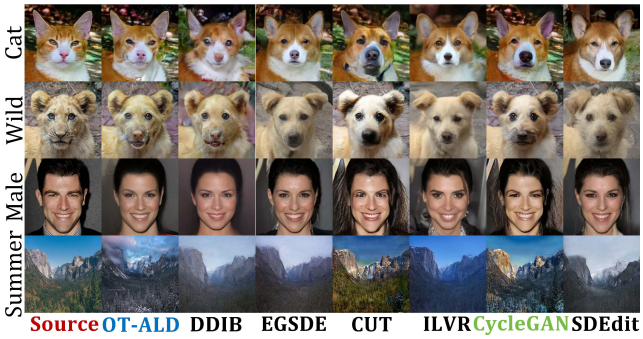



Figure 6: Visualization comparison of different models over four I2I translation tasks. Compared with other mainstream methods, OT-ALD has the highest fidelity to source images, and its output quality is on par with or even superior to them.

FID value of our model is 2.6 lower on average than that of the optimal model in each corresponding task. Furthermore, Figure 6 provides a visual comparison between OT-ALD and representative I2I models, showing that our model generates clearer and more realistic results than its counterparts, while largely maintaining structure of the source image.

Human evaluations offer a more accurate assessment of I2I translation quality. As shown in (Mantiuk, Tomaszewska, and Mantiuk 2012), forced pairwise comparison is a reliable method for image quality evaluation. Accordingly, we adopt the 2-alternative forced-choice paradigm (Zhang, Isola, and

| Methods | Cat→Dog | | | |
|---|---|---|---|---|
| | FID↓ | $L_2$ ↓ | PSNR↑ | SSIM↑ |
| CUT | 76.21 | 59.78 | 17.48 | **0.601** |
| StarGANv2 | 54.88 | 133.65 | 10.63 | 0.27 |
| CycleGAN | 85.90 | — | 18.13 | 0.58 |
| QS-Attn | 72.80 | — | 18.26 | 0.39 |
| SDEdit | 74.14 | 47.88 | 19.19 | 0.423 |
| SDDM | 62.29 | — | — | 0.422 |
| SDDM† | 49.43 | — | — | 0.361 |
| InjectDiff | 63.76 | — | 20.10 | 0.48 |
| ILVR | 74.37 | 56.95 | 17.77 | 0.363 |
| EGSDE | 65.82 | 47.22 | 19.31 | 0.415 |
| EGSDE† | 51.04 | 62.02 | 17.17 | 0.361 |
| DMT | 47.86 | 45.98 | 20.55 | 0.412 |
| DDIB | 59.62 | 60.23 | 18.91 | 0.372 |
| OT-ALD | **44.31** | **42.35** | **21.20** | 0.470 |

| Methods | Wild→Dog | | | |
|---|---|---|---|---|
| | FID↓ | $L_2$ ↓ | PSNR↑ | SSIM↑ |
| CUT | 92.94 | 62.21 | 17.20 | **0.592** |
| ILVR | 75.33 | 63.40 | 16.85 | 0.287 |
| SDEdit | 68.51 | 55.36 | 17.98 | 0.343 |
| EGSDE | 59.75 | 54.34 | 18.14 | 0.343 |
| EGSDE† | 50.43 | 66.52 | 16.40 | 0.300 |
| SDDM† | 57.38 | — | — | 0.328 |
| DMT | 49.83 | 52.42 | 18.05 | 0.338 |
| DDIB | 55.34 | 60.01 | 17.72 | 0.332 |
| OT-ALD | **47.65** | **50.03** | **18.95** | 0.351 |

| Methods | Male→Female | | | |
|---|---|---|---|---|
| | FID↓ | $L_2$ ↓ | PSNR↑ | SSIM↑ |
| CUT | 31.94 | 46.61 | 19.87 | 0.74 |
| ILVR | 46.12 | 52.17 | 18.59 | 0.510 |
| SDEdit | 49.43 | 43.70 | 20.03 | 0.572 |
| EGSDE | 41.93 | 42.04 | 20.35 | 0.574 |
| EGSDE† | 30.61 | 53.44 | 18.32 | 0.510 |
| InjectDiff | 28.12 | — | 21.65 | 0.67 |
| CycleGAN | 36.75 | — | 21.54 | 0.70 |
| QS-Attn | 32.56 | — | 20.68 | 0.60 |
| SDDM | 44.37 | — | — | 0.526 |
| DMT | 29.01 | 35.63 | 22.42 | 0.687 |
| DDIB | 38.25 | 38.55 | 21.37 | 0.66 |
| OT-ALD | **25.21** | **31.51** | **23.05** | **0.75** |

| Methods | Summer→Winter | | |
|---|---|---|---|
| | FID↓ | KID↓ | DINO Struct.↓ |
| CycleGAN | 62.9 | 1.022 | 2.6 |
| DistanceGAN | 97.2 | 2.843 | — |
| GcGAN | 97.5 | 2.755 | — |
| CUT | 72.1 | 1.207 | 2.1 |
| SDEdit | 66.1 | 3.218 | 2.1 |
| Plug&Play | 67.3 | — | 2.8 |
| CycleDiff | 64.1 | — | 3.6 |
| P2P | 99.1 | 2.626 | — |
| Pix2Pix-Zero | 68.0 | — | 3.0 |
| InstPix2Pix | 68.3 | — | 3.7 |
| UNSB | 73.9 | **0.421** | — |
| DDIB | 90.8 | 2.36 | 7.2 |
| Turbo | 56.3 | — | **0.6** |
| NiceGAN | 76.03 | 0.67 | — |
| NiceGAN† | 77.13 | 0.73 | — |
| U-GAT-IT | 88.41 | 1.43 | — |
| EGSDE | 62.38 | 0.75 | 1.8 |
| ILVR | 65.64 | 1.35 | 2.0 |
| DMT | 54.64 | 0.564 | 1.7 |
| OT-ALD | **52.87** | 0.521 | 1.5 |

Table 1: Quantitative comparison on four tasks. ↑ and ↓ indicate directions of better values, where the optimal value is marked in **bold** and the suboptimal is marked with underline.

Efros 2016), where 50 participants are shown 20 randomly selected image pairs, each consisting of a generated image and its corresponding real target image. Participants are informed that only one image per pair is real and asked to choose the one they believe is real. The proportion of a generated image is mistakenly selected, *Fool Rate*, is recorded. The Fool Rate closer to 50% indicates higher image realism, while a lower suggests more noticeable artifacts.

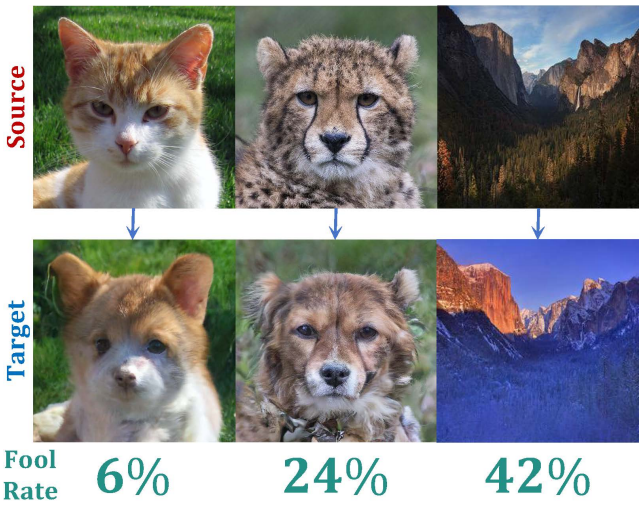



Figure 7: Visualization comparison of Fool Rate. The Fool Rate essentially reflects the difficulty in distinguishing between translated and real images. The larger Fool Rate is and the closer it is to 50%, the stronger the realism of the images.

Figure 7 shows examples generated by OT-ALD with their corresponding average Fool Rates. Some results from the summer-to-winter task achieve rates above 40%, indicating high realism, while some distorted images score near 0%. Table 2 compares the average Fool Rates of our model with others across four tasks, where our model consistently achieves around 29%, suggesting it often produces images that are partially indistinguishable from real ones.

## Translation Efficiency

To evaluate the computational efficiency of different models on various translation tasks, we first fully train them on a single RTX 4090 GPU. Then, for each task, we randomly select 1000 source images and measure the average GPU time required by each model to translate them into target images. The results in Table 3 demonstrate that OT-ALD is more efficient than other diffusion-based I2I translation methods under same conditions. This advantage mainly stems from the reduced $T$ required by our model.

In addition to keeping the experimental equipment consistent, when testing the translating time in Table 3, we adopted the same sampling method for all diffusion-based models and did not use any acceleration strategies for the diffusion process. This means that if an acceleration strategy is introduced, all time data in Table 3 will decrease synchronously, while our model will still outperform other baseline models.

| Methods | Cat→Dog | Wild→Dog | Summer→Winter | Male→Female |
|---|---|---|---|---|
| CUT | 23.6% | 22.3% | 29.9% | 28.5% |
| CycleGAN | 22.3% | 25.4% | 34.5% | 23.7% |
| ILVR | 21.4% | 24.1% | 23.3% | 29.2% |
| EGSDE | 24.1% | **28.3%** | 26.6% | 27.2% |
| DDIB | 18.5% | 19.9% | 30.8% | 25.9% |
| SDEdit | 20.7% | 21.5% | 33.4% | 24.4% |
| DMT | **24.3%** | 25.7% | 33.9% | 27.5% |
| OT-ALD | 22.8% | 26.9% | **37.2%** | **29.4%** |

Table 2: Comparison of average Fool Rates (↑) of different methods.

| Methods | Cat→Dog | Wild→Dog | Summer→Winter | Male→Female |
|---|---|---|---|---|
| EGSDE | 79.85 (51.04) | 79.75 (40.43) | 19.06 (62.38) | 79.80 (30.61) |
| ILVR | 52.31 (70.37) | 52.26 (75.33) | 16.97 (65.64) | 52.58 (46.12) |
| DDIB | 92.30 (59.62) | 92.36 (55.34) | 20.64 (90.80) | 92.35 (38.25) |
| DMT | 75.20 (47.86) | 74.98 (49.83) | 18.85 (54.64) | 75.25 (29.01) |
| SDEdit | 86.44 (74.14) | 87.12 (68.51) | 19.85 (66.10) | 86.77 (49.43) |
| OT-ALD | **45.15 (44.31)** | **45.18 (47.65)** | **10.21 (52.87)** | **45.18 (25.21)** |

Table 3: Translating time comparison (s/image, ↓) of different *diffusion-based* models *under same conditions*. The corresponding FID scores (↓) are indicated in parentheses. Our model adopts $T = 500$ and $\eta = 0.2$. The average sampling time of our model across four tasks is 20.29% lower than that of the model with the best translation efficiency (ILVR).

## Ablation Study

(1) **Optimal Transport.** Figures 1(b)-1(c) and Figure 5 compare the DDIB-based method and OT-ALD in terms of latent distribution alignment, translated target distribution, cycle consistency, and their variations over diffusion time $T$.

(2) **Noise Scale Factor $\eta$.** Figure 3(a) illustrates the impact of $\eta$ on OT-ALD. With fixed $T$, a larger $\eta$ leads to improved output quality and diversity.

(3) **Diffusion termination time $T$.** Figure 3(b) demonstrates the effect of $T$ on OT-ALD. When $\eta$ is fixed and positive, increasing $T$ enhances both generation quality and diversity. Moreover, for fixed-size image tasks, the average translation time of OT-ALD increases linearly with $T$.

## Conclusion

This paper provides a theoretical analysis of the limitations of classical DDIB-based I2I translation methods, namely: (1) low efficiency during the translation process and (2) translation trajectory deviations due to mismatched latent distributions. To address these issues, we propose OT-ALD, which effectively eliminates latent distribution mismatch, and offers theoretical guarantees for key properties such as cycle consistency. In practice, OT-ALD enables fast, high-quality image translation with flexible control over transformation objectives, allowing users to balance generation speed and synthesis quality according to their specific needs.

## Acknowledgments

This research was supported by the National Key Research and Development Program of China under Grant No. 2021YFA1003003; the National Natural Science Foundation of China under Grant No. T2225012.

# References


Anderson, B. D. 1982. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3): 313–326.

Azadi, S.; Fisher, M.; Kim, V. G.; Wang, Z.; Shechtman, E.; and Darrell, T. 2018. Multi-content gan for few-shot font style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7564–7573.

Benaim, S.; and Wolf, L. 2017. One-sided unsupervised domain mapping. *Advances in neural information processing systems*, 30.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Bourou, A.; Boyer, T.; Gheisari, M.; Daupin, K.; Dubreuil, V.; De Thonel, A.; Mezger, V.; and Genovesio, A. 2024. PhenDiff: Revealing subtle phenotypes with diffusion models in real images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 358–367. Springer.

Brenier, Y. 1991. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44: 375–417.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402.

Chen, R.; Huang, W.; Huang, B.; Sun, F.; and Fang, B. 2020. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8168–8177.

Cheng, B.; Liu, Z.; Peng, Y.; and Lin, Y. 2023. General image-to-image translation with one-shot image guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22736–22746.

Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8188–8197.

Franzese, G.; Rossi, S.; Yang, L.; Finamore, A.; Rossi, D.; Filippone, M.; and Michiardi, P. 2023. How much is enough? a study on diffusion times in score-based generative models. *Entropy*, 25(4): 633.

Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Zhang, K.; and Tao, D. 2019. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2427–2436.

Gu, X.; Luo, F.; Sun, J.; and Yau, S.-T. 2013. Variational principles for Minkowski type problems, discrete optimal transport, and discrete Monge-Ampere equations. *arXiv preprint arXiv:1302.5472*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6629–6640.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 6840–6851.

Hu, X.; Zhou, X.; Huang, Q.; Shi, Z.; Sun, L.; and Li, Q. 2022. Qs-attn: Query-selected attention for contrastive learning in i2i translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18291–18300.

Huang, Y.; Huang, J.; Liu, Y.; Yan, M.; Lv, J.; Liu, J.; Xiong, W.; Zhang, H.; Chen, S.; and Cao, L. 2024. Diffusion model-based image editing: A survey. *arXiv preprint arXiv:2402.17525*.

Hur, J.; Choi, J.; Han, G.; Lee, D.-J.; and Kim, J. 2024. Expanding expressiveness of diffusion models with limited data via self-distillation based fine-tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5028–5037.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jiang, H.; Imran, M.; Ma, L.; Zhang, T.; Zhou, Y.; Liang, M.; Gong, K.; and Shao, W. 2024. Fast-DDPM: Fast Denoising Diffusion Probabilistic Models for Medical Image-to-Image Generation. *arXiv e-prints*, arXiv–2405.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Khrulkov, V.; Ryzhakov, G.; Chertkov, A.; and Oseledets, I. 2022. Understanding DDPM Latent Codes Through Optimal Transport. In *The Eleventh International Conference on Learning Representations*.

Kim, B.; Kwon, G.; Kim, K.; and Ye, J. C. 2023. Unpaired Image-to-Image Translation via Neural Schr\" odinger Bridge. *arXiv preprint arXiv:2305.15086*.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2426–2435.

Kim, J.; Kim, M.; Kang, H.; and Lee, K. 2019. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*.

Kwon, G.; and Ye, J. C. 2022. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*.

Lee, E.; Jeong, S.; and Sohn, K. 2024. EBDM: Exemplar-Guided Image Translation with Brownian-Bridge Diffusion Models. In *European Conference on Computer Vision*, 306–323. Springer.

Lee, J.; Kang, M.; and Han, B. 2024. Diffusion-Based Image-to-Image Translation by Noise Correction via Prompt Interpolation. In *European Conference on Computer Vision*, 289–304. Springer.

Lei, N.; An, D.; Guo, Y.; Su, K.; Liu, S.; Luo, Z.; Yau, S.-T.; and Gu, X. 2020. A geometric understanding of deep learning. *Engineering*, 6: 361–374.

Li, B.; Xue, K.; Liu, B.; and Lai, Y.-K. 2023. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, 1952–1961.

Li, L.; and Ma, L. 2023. Injecting-Diffusion: Inject Domain-Independent Contents into Diffusion Models for Unpaired Image-to-Image Translation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 282–287. IEEE.

Luo, S.; Tan, Y.; Huang, L.; Li, J.; and Zhao, H. 2023. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*.

Mancusi, M.; Halychanskyi, Y.; Cheuk, K. W.; Moliner, E.; Lai, C.-H.; Uhlich, S.; Koo, J.; Martínez-Ramírez, M. A.; Liao, W.-H.; Fabbro, G.; et al. 2024. Latent Diffusion Bridges for Unsupervised Musical Audio Timbre Transfer. *arXiv preprint arXiv:2409.06096*.

Mantiuk, R. K.; Tomaszewska, A.; and Mantiuk, R. 2012. Comparison of four subjective methods for image quality assessment. In *Computer graphics forum*, volume 31, 2478–2491. Wiley Online Library.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 319–345. Springer.

Parmar, G.; Kumar Singh, K.; Zhang, R.; Li, Y.; Lu, J.; and Zhu, J.-Y. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

Parmar, G.; Park, T.; Narasimhan, S.; and Zhu, J.-Y. 2024. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, 1–10.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022b. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4713–4726.

Sasaki, H.; Willcocks, C. G.; and Breckon, T. P. 2021. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*.

Seo, J.; Lee, G.; Cho, S.; Lee, J.; and Kim, S. 2023. Midms: Matching interleaved diffusion models for exemplar-based image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2191–2199.

Shi, Y.; Xue, C.; Liew, J. H.; Pan, J.; Yan, H.; Zhang, W.; Tan, V. Y.; and Bai, S. 2024. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8839–8849.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, Y.; Durkan, C.; Murray, I.; and Ermon, S. 2021. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34: 1415–1428.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Su, X.; Song, J.; Meng, C.; and Ermon, S. 2022. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*.

Sun, S.; Wei, L.; Xing, J.; Jia, J.; and Tian, Q. 2023. SDDM: score-decomposed diffusion models on manifolds for unpaired image-to-image translation. In *International Conference on Machine Learning*, 33115–33134. PMLR.

Tumanyan, N.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2022. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10748–10757.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1921–1930.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, C. H.; and De la Torre, F. 2023. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7378–7387.

Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; Timotfe, R.; and Van Gool, L. 2024a. Diffi2i: efficient diffusion model for image-to-image translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xia, M.; Zhou, Y.; Yi, R.; Liu, Y.-J.; and Wang, W. 2024b. A diffusion model translator for efficient image-to-image translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18381–18391.

Yi, Z.; Zhang, H.; Tan, P.; and Gong, M. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, 2849–2857.

Yin, W.; Yu, Y.; Yin, H.; Kragic, D.; and Björkman, M. 2024. Scalable motion style transfer with constrained diffusion generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10234–10242.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

Zhang, J.; Rimchala, J.; Mouatadid, L.; Das, K.; and Kumar, S. 2024. DECDM: Document enhancement using cycle-consistent diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 8036–8045.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, 649–666. Springer.

Zhao, M.; Bao, F.; Li, C.; and Zhu, J. 2022. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *Advances in Neural Information Processing Systems*, 35: 3609–3623.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.